ARMY RESEARCH LABORATORY

Adding Statistical Machine Translation Adaptation to Computer-Assisted Translation

by Robert P. Winkler, Somiya Metu, Stephen A. LaRocca, and John J. Morgan

ARL-TR-6637 **September 2013**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-TR-6637 **September 2013**

Adding Statistical Machine Translation Adaptation to Computer-Assisted Translation

Robert P. Winkler, Somiya Metu, Stephen A. LaRocca, and John J. Morgan
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
September 2013	Final	

4. TITLE AND SUBTITLE
Adding Statistical Machine Translation Adaptation to Computer-Assisted Translation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Robert P. Winkler, Somiya Metu, Stephen A. LaRocca, and John J. Morgan

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CII-B
2800 Powder Mill Road
Adelphi, MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-6637

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Statistical machine translation (SMT) has proven effective for general purpose language translation but not for highly specialized domains like medicine, military operations, and law enforcement, which have their own technical jargon. We present a novel approach for iteratively incorporating a human translator in the loop to adapt SMT models to a particular domain. We show how these models can be made accessible via Web services and integrated with computer-assisted translation (CAT) tools. In this report, we describe a novel human-in-the-loop post-editing domain adaptation algorithm for refining SMT models using the Joshua decoder and integrate it with a CAT tool called OmegaT.

15. SUBJECT TERMS
Statistical machine translation, computer-assisted translation, domain adaptation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Robert P. Winkler
Unclassified	Unclassified	Unclassified	UU	22	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-1689

Contents

List of Figures

List of Listings

1. Introduction

Computer-assisted translation (CAT) tools provide human translators with a software system to support and facilitate their translation activities. Typical functionality includes (1) Translation Memory (TM), which contains previously translated segments that are automatically searched and used to suggest possible translations; (2) spell-checkers; (3) glossaries; (4) dictionaries; (5) alignment and segmentation tools allowing the translator to decide the segment boundaries (typically sentence or paragraph); and most recently, (6) machine translation (MT) support.

Statistical machine translation (SMT) is a paradigm for MT based on statistical analyses of parallel bilingual text data. The U.S. Army Research Laboratory (ARL) has developed a technique for SMT model adaptation and applied it to a variety of specialized domains (medical, legal, military, agriculture, etc.) to assist English to Dari document translation in support of Operation Enduring Freedom (OEF) in Afghanistan. Unlike traditional SMTs, which typically seek to be as general purpose as possible, many of these domains have highly specialized and peculiar jargon, which is unlikely to be captured in a general purpose model. Further complicating the translation process is the dearth of translators with bilingual subject matter expertise for the low resource target languages of Army transition and training operations in Afghanistan. In the course of site visits to operational units in theater, researchers at ARL perceived that the current model of hiring a roomful of (expensive) translators to translate individual documents is wasteful, in that it fails to methodically capture the expertise of these translators. While translated documents are the end product of language operations, capturing the translators' expertise in an automated fashion can serve to bootstrap the translation of future documents in the same domain. Moreover, it can promote consistent use of technical terms in the target languages and increase the overall effectiveness of human translators, especially as they rotate in and out of their job, as happens frequently in Afghanistan and throughout the U.S. Central Command (CENTCOM). ARL researchers developed a process we call "iterative post-editing domain adaptation for SMT" to capture the translators' expertise. Central to this process is both the careful management of how bilingual subject matter experts perform their translation work, so that carefully aligned parallel sentences result, and a very short software development cycle, which presents the human experts with draft translations incorporating the same choices they made in their most recent translation assignments. As SMT models are recomputed with each new chapter or "chunk" of text, they become more tightly focused on the domain of interest.

Once these highly specialized, domain-specific SMT models have been created, the next step is to determine how to effectively use them outside the laboratory environment in which they were developed. To that end, we created a Web service for each SMT model, or SMT engine, to

accommodate a variety of different front-ends from browser-based thin clients to workstation applications. We realized that, of the various ways that government-owned Web translation services might be offered to Army users, integration with existing CAT tools might be particularly useful, as the same CAT system can readily provide access to other translator tools, such as bilingual glossaries and spell-checkers.

This report describes the iterative post-editing domain adaptation algorithm, as well as a method by which domain-specific SMT can be offered as Web services and a procedure for the integration of Web-based translation services with an open-source CAT tool called OmegaT.

2. Iterative Post-Editing Domain Adaptation for SMT

The current model in the Army for the translation of low resource languages (e.g., Dari) for which there are few expert translators is to form teams of bilingual contractors who work together under a supervising staff. With today's emphasis on military training and security assistance operations, such teams concentrate on translating specific types of documents with the aim of generating high-quality translations. The tools these teams typically have available are word processors and access to some existing translation resources on the Web. Some well-led teams share glossaries of parallel aligned terms (usually in the form of Excel spreadsheets), but other resources to help automate the translation process are typically not available. To further complicate the translation work, many of the documents identified are in a highly specialized domain (medicine, legal, agricultural, etc.) with terms that may not be familiar to the translators. ARL researchers noted that this current model focuses almost exclusively on the translated document itself as the end product, to the detriment of translation as a process. As a result, the expert knowledge that is generated during the course of the translation is preserved only in the mind of the human translators. When human translators finish their assignments, valuable knowledge leaves with them.

There are a number of artifacts generated during the translation effort, which with the help of technology can be preserved to further assist future translations within the highly specialized domain. One method for preserving expert translation knowledge is to use automatic term extraction to identify domain-specific language in a translated document. A variant of tf-idf (Luhn, 1958; Yu and Salton, 1976; Robertson and Sparck Jones, 1976) can be used to identify the most important phrases and terms to generate a glossary for the domain. Another method is to create parallel text from the archives of a team's translation work and align that text at the sentence level; such a corpus can then be used to produce SMT models adapted to the domain in which the team is specialized. In turn, this SMT can be used to generate initial draft translations, or translations hypotheses, for successive documents. When enough parallel data is available for

a mature SMT model, the translator just edits the translation hypotheses instead of starting a new translation completely from scratch.

Our approach to capturing this expert knowledge is to think of translation work in terms of a project, each of which deserves its own SMT model. We first start with a baseline SMT model trained on the best parallel data available and use that to generate hypotheses for a portion (we say "chunk") of the entire translation project. For the baseline SMT model used in our legal Dari SMT project, ARL had collected some 46,000 parallel English/Dari sentences, only some of which pertained to the legal domain. These were used to produce a baseline SMT model with which the first chunk of the translation project was translated, i.e., rendered as a set of hypotheses. According to our approach, these translation hypotheses are then given to a human translator to correct, or post-edit. When the human is finished with that portion of the document, the results are used to retrain the SMT model and retune it for better accuracy on the domain of interest. This resulting new SMT model is then used to generate hypotheses for the next portion of the document. This process continues until the document comprising the project has been completely translated with the end result being an SMT model highly tuned to the specific domain (as well as to the specific human translator and the specific document). While the process has been shown to produce high quality translations with excellent consistency of terminology, it raises the concern of overfitting the SMT model to a very narrow purpose and to the preferences of a single translator.

Happily, the continuous nature of the SMT development approach means the final SMT model for one project is only the initial, or baseline, SMT model for the next one. As more and more documents are translated by different human translators, the effect of any single document and any single translator diminishes and the SMT model effectively becomes more representative of the team operating in their habitual domain.

A formal statement of the algorithm is shown in listing 1 and the data flow diagram is shown in figure 1.

Definitions

Let a segment be a sequence of terms. Let *s* be a segment in a source language and *t* a segment in a target language. Let an ordered pair *(s,t)* represent a translation pair.

Then a bilingual parallel corpus $C := \{(s_i,t_i): 0 \le i < |C|, s_i \neq \emptyset, t_i \neq \emptyset\}$ where t_i is assumed to be an expert translation of segment s_i from the source to target language.

Let $TRAIN_i := \{(s_j,t_j): 0 \le j < |TRAIN_i|, s_j \neq \emptyset, t_j \neq \emptyset\}$ and $TUNE_i := \{(s_j,t_j): 0 \le j < |TUNE_i|, s_j \neq \emptyset, t_j \neq \emptyset\}$ where t_i is assumed to be an expert translation of segment s_j from the source to target language be the i^{th} iteration training and tuning sets of translation, respectively.

Let $TERMS_i := \{term: \forall(s_j \in \cap\cup TRAIN_i)\ term \in s_j\}$ be the set of all source language terms in the i^{th} iteration training set.

Let $TM_i(TRAIN_i, TUNE_i)$:= the translation model resulting from the i^{th} iteration training and tuning sets.

Let D := the document to be translated and $D_i \subset D$:= the i^{th} set of source segments to be translated.

Let a translation hypotheses $H_i := \{(s_j, t_j): s_j \in D_i\}$ where t_j represents a proposed translation of segment s_j from the source to target language. Note that if any term in s_j is out of vocabulary (i.e. it doesn't appear anywhere in $TERMS_i$), t_j will be a mix of the source and target language and if all of the terms in s_j are out of vocabulary, $t_j = s_j$.

Let a translation $T_i := \{(s_j, t_j): s_j \neq \emptyset, t_j \neq \emptyset, \forall(s_j) s_j \in \cap\cup H_i\}$ be the post-edited human-generated translation of H_i.

Finally let $OOV_i := \{(s_j, t_j): s_j \in \cap\cup T_i, \exists(term \in s_j): term \notin TERMS_i\}$ be the set of all source segments in T_i where there is at least one term in s_j which does not appear in $TERMS_i$.

Initial Conditions

$TRAIN_0$ and $TUNE_0$ are nonempty disjoint proper subsets of C the union of which is C: $TRAIN_0 \subset C, TUNE_0 \subset C, TRAIN_0 \cup TUNE_0 = C, TRAIN_0 \cap TUNE_0 = \emptyset, TRAIN_0 \neq \emptyset, TUNE_0 \neq \emptyset$ and the size of $TUNE_0$ is much less than the size of $TRAIN_0$: $|TUNE_0| \ll |TRAIN_0|$.

Algorithm

```
i=0
while D ≠ Ø
        // train and tune the i^th translation model
        TM_i(TRAIN_i, TUNE_i)
        // select the i^th subset of segments left in D to translate
        Select D_i
        // Update D
        D = D – D_i
        // Generate the i^th translation hypotheses by decoding D_i
        H_i = Decode(TM_i, D_i)
        // Give the hypothesis to the translator
        T_i = Postedit(H_i)
        // Update the i^th+1 training set with all the out of vocabulary translation pairs
        TRAIN_{i+1} = TRAIN_i ∪ OOV_i
        // Update T_i
        T_i = T_i – OOV_i
        // Split the reminder of T_i between the i^th+1 training and tuning sets
        { NewTrain, NewTune } = Split(T_i)
        // Update the i^th+1training and tuning sets
        TRAIN_{i+1} = TRAIN_i ∪ NewTrain
        TUNE_{i+1} = TUNE_i ∪ NewTune
        i = i+1
done
```

Listing 1. Iterative post-editing domain adaptation for SMT algorithm.

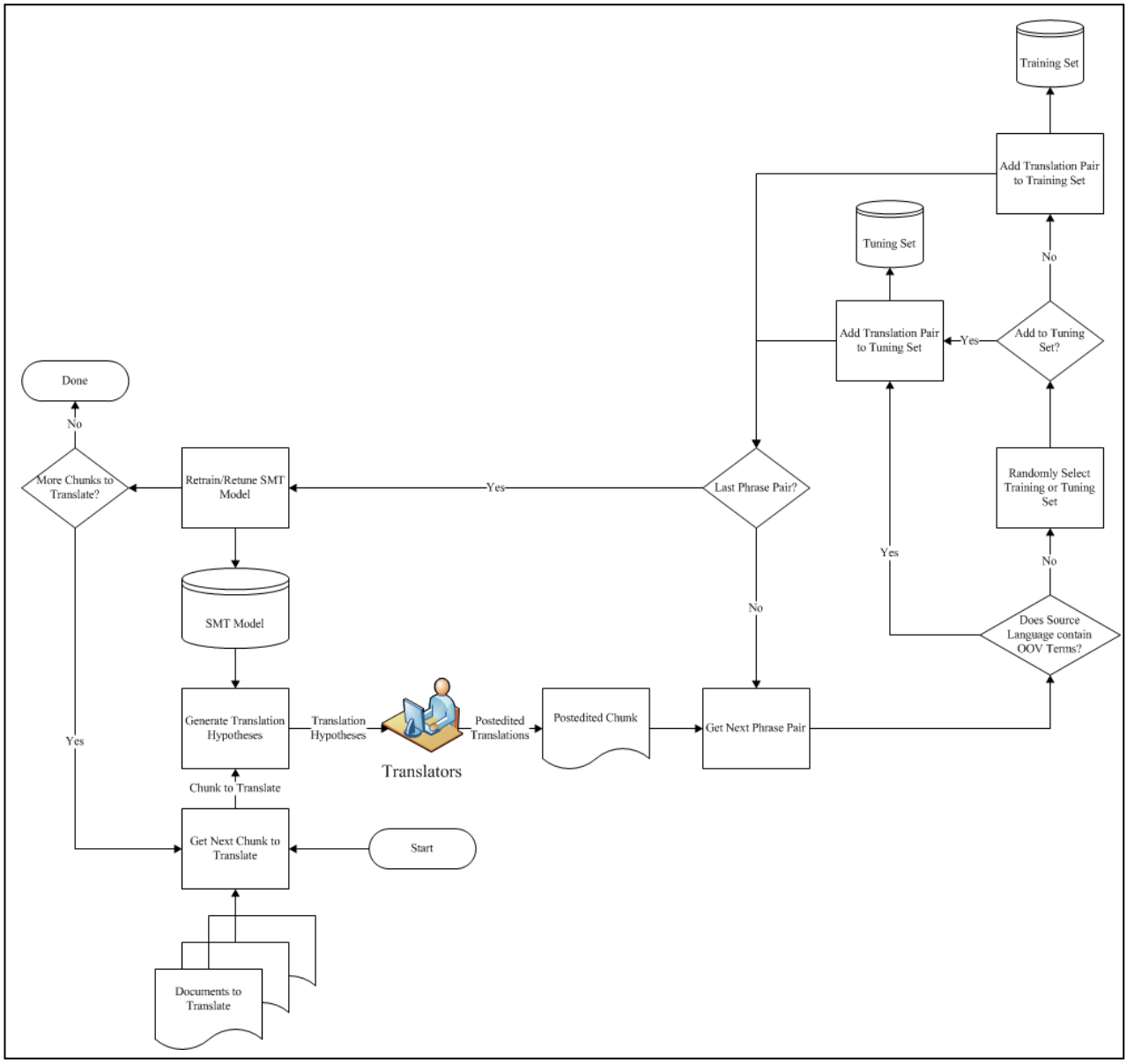

Figure 1. Iterative post-editing domain adaptation data flow.

3. Related Work

In the field of CAT, most modern CAT tools already provide some level of access to MT and some work has been done to quantify its impact on interactive translation (Koehn, 2009; Plitt and Messelot, 2010), but these systems typically do not provide MT support for low resource languages like Dari. Incorporating MT for new language pairs or tailoring an existing model to a specific domain in proprietary CAT tools, as with SMT domain adaptation, also incurs additional

costs since the MT vendor is usually involved. Government-owned SMT models make with open-source SMT systems like Joshua and open-source CAT tools like OmegaT avoid these costs while providing versatility for translation work specific to domains of military interest.

Other researchers have looked at applying SMT as a "statistical post-editing" step to improve the quality of output from rule-based MT systems like SYSTRAN. Their focus is on combining the two MT paradigms to boost overall accuracy, while ours instead is on repeatedly using human-in-the-loop translators to further refine and adapt SMT models within the context of a single paradigm (Dugast, Senellart, and Koehn, 2007; Simard et al., 2007; Terumasa, 2007; Lagarda et al., 2009).

Some related work has also been done on using active learning for SMT. Most of the work focuses on optimizing sentence selection to give to the human translators (Callison-Burch, 2003; Mandal et al., 2008; Haffari and Sarkar, 2009; Ambati and Vogel, 2010; González-Rubio, Ortiz-Martínez, and Casacuberta, 2011; Ananthakrishnan et al., 2011; Bakhshaei and Khadivi, 2012). In our work, sentence selection is driven by their order in the document. To generate the highest quality translation, the translators are given the sentences in context. Other related work on SMT domain adaptation include the use of mixture models and classifiers and metrics to distinguish a particular domain and/or incorporate multiple domain models in a single system (Bertoldi and Federico, 2009; Foster, Goutte, and Kuhn, 2010; Wang et al., 2012; Sennrichm 2012). In our work, the focus is on improving the quality of the translation and the experience of the translators where the domain and SMT model is known beforehand. While some commercial SMT systems like SDL Language Weaver (http://www.sdl.com/products/sdl-enterprise-language-server/) have provisions to generate SMT models from user data for highly specialized domains similar to our iterative post-editing domain adaptation algorithm, such customized trainings incur significant software licensing costs and annual software maintenance costs.

Morgan (2010) first used this iterative post-editing domain adaptation algorithm to assist in the translation of the medical training manual *Fundamental Critical Care Support*, published by the Society for Critical Care Medicine, into Dari. He used it again in 2011 to translate U.S. Army Field Manual 7-8 (*The Rifle Platoon and Squad*) from English into Pashto (Morgan, 2011). In his reports, he notes increased translation accuracy as measured in terms of rising Bilingual Evaluation Understudy (BLEU) scores as each chapter or chunk of the manual is automatically rendered as a draft translation, corrected (post-edited) by a human expert, and then used to retrain and retune the SMT model before it is used to produce a draft translation of the next chapter or chunk. The work here builds on Morgan's work but is more selective in which of the post-edited translations go into the training set and which go into the tuning set. Instead of retraining and retuning on the entire chunk of post-edited text, any source language sentences that contain out of vocabulary (OOV) terms go into the training set along with their translation, so the SMT model has at least one translation for the OOV item. The remaining parallel text is

split between the training and tuning sets by a user-defined parameter (we used 50% in training and 50% in tuning).

4. Web Service Front-End to the Joshua SMT System

Once we have a suite of refined, highly tuned, domain-specific SMT models, we want to expose them for use by other applications. The SMT decoder we used for this effort is the open-source Joshua decoder developed by the Center for Language and Speech Processing and the Human Language Technology Center of Excellence at the Johns Hopkins University (http://joshua-decoder.org/). Proprietary SMT decoders currently in use by the Army come with a high cost. Some systems have license costs greater than $100,000 with as much as a 15% annual maintenance fee. The Joshua decoder is free and provides a mechanism for individual researchers and users to retrain and retune the SMT on different domains. In a research environment, the standard stable release of the Joshua decoder is typically invoked via the command line. A user sends a list of source language sentences and receives translation hypotheses back. An issue with this command-line interface is that the entire SMT engine is reinitialized for every list of sentences, which is inefficient and results in long latencies. The development branch of the Joshua decoder, however, contains a client-server interface where the client application sends the sentence(s) over a TCP/IP socket and the server sends the translation hypotheses back without having to reinitialize the system each time. While this efficiently and effectively exposes the SMT models, it requires the client application to reside on the same network where the server lives. To expose the SMT models to a wider community, we wrapped the Joshua decoder server with a simple Representational State Transfer (REST) Web service. For each domain, we create a different instance of the Joshua decoder each with a different SMT model and URL for the Web service as shown in figure 2.

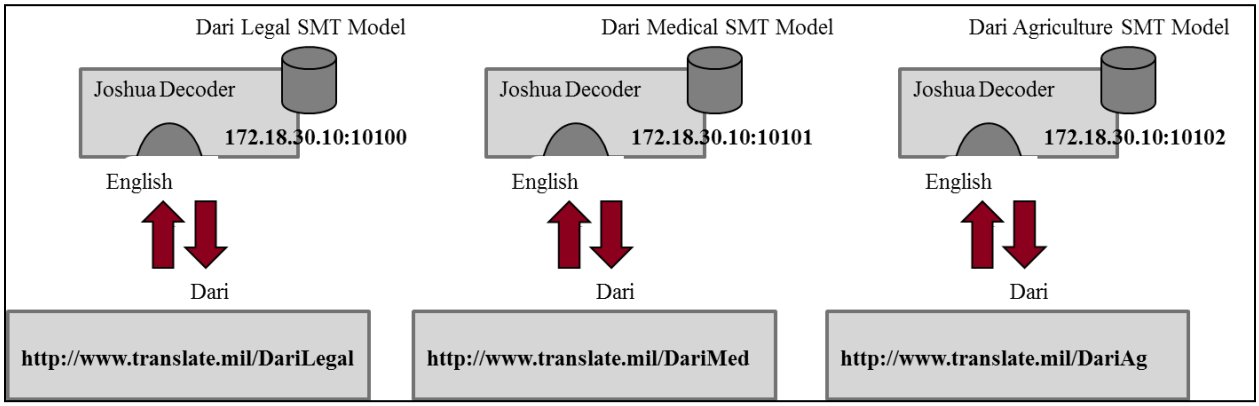

Figure 2. Deployment of various domain-specific SMT models.

The Web service itself is extremely lightweight and simple. Clients HTTP POST one or more English sentences to the URL and the HTTP Response contains the translations. A typical translation transaction is shown in listings 2 and 3.

```
POST http://www.translate.mil/DariLegal HTTP/1.1
Fingerprints, although they may be found 50 years after being deposited on a piece of paper, are at the same time very fragile and easily destroyed.
```

Listing 2. Example English HTTP POST.

```
HTTP/1.1 200 OK

کشف انگشت ، با آن ممکن است بعد از گذشت ۵۰ سال هم بعد از تشکل بر روی یک پارچه کاغذ قابل بازیابی هستند ، در عین حال بسیار شکننده بوده و به آسانی تخریب گردیده است .
```

Listing 3. Example Dari response.

When a client makes a connection and submits the English sentences, the Web service first stores the sentences in a temporary file. Since the Joshua decoder works best when the sentences are tokenized and lower case, a pipeline of Perl scripts is invoked. First, a script to tokenize the sentences is executed; the result is then sent to another script, which converts it to lowercase and then the final result is sent to the Joshua decoder server via the Netcat utility and the Web service captures the result from stdout. The data flow is shown in figure 3 and the shell script in listing 4.

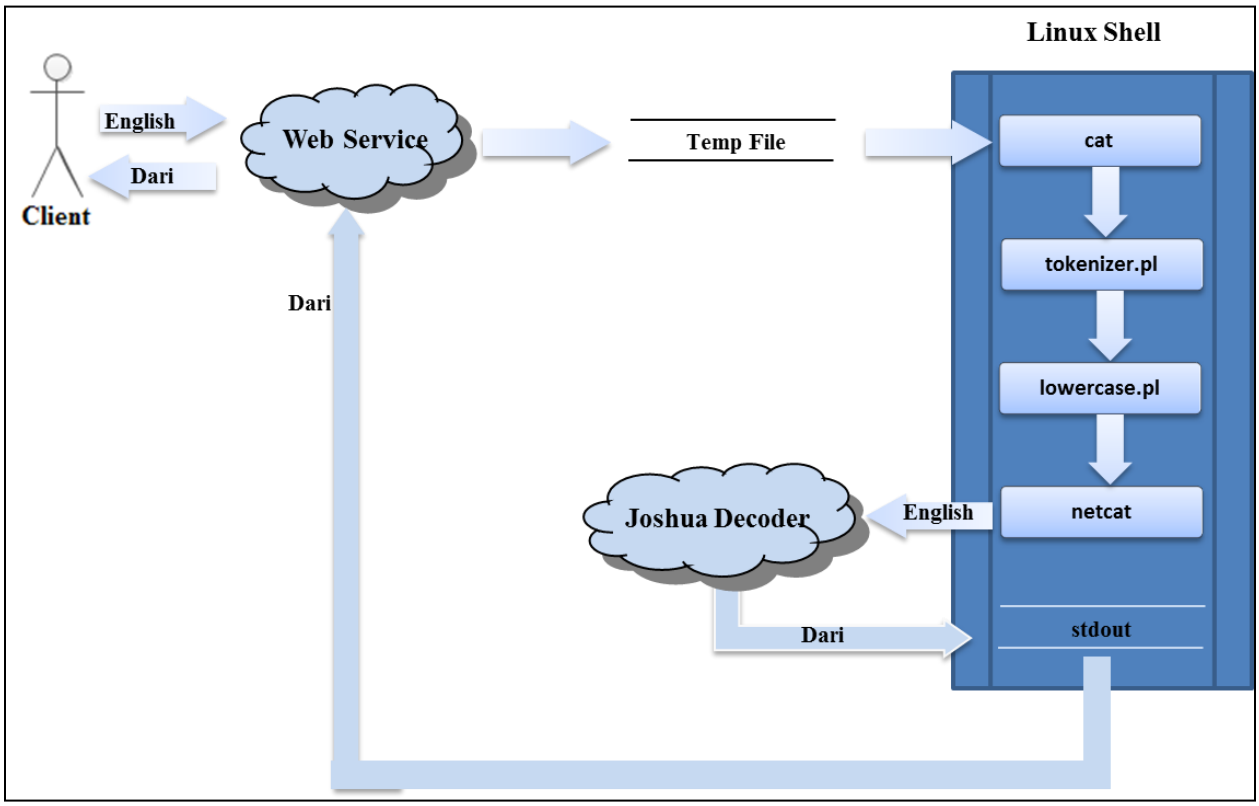

Figure 3. English-to-Dari translation dataflow.

```
cat EnglishTemp.txt | perl tokenizer.pl | perl lowercase.pl | nc localhost 10101
```

Listing 4. English-to-Dari translation shell script.

5. Integration with OmegaT

OmegaT is a free CAT tool intended for professional translators. Its list of features include creation and management of TM; TM search capabilities; fuzzy matching against TMs to propose translations; spell-checking, glossary, and dictionary look-up; support for multiple file formats; regular expressions; programmable segmentation; and auto replacement of suggested translations (http://www.omegat.org/en/omegat.html).

OmegaT was selected as the CAT tool to integrate with because it's open source and extensible. OmegaT is written in the Java language with extensibility provided via interfaces and abstract base classes. Prepackaged MT plug-ins are provided for translations from Apertium (http://www.apertium.org), Belazar, and Google Translate (http://translate.google.com/). To facilitate the incorporation of additional MT systems, OmegaT defines an interface

IMachineTranslation, which specifies two functions for getting the name of the translation plugin (*getName*()) and the translation itself (*getTranslation*(…)), and an abstract base class *BaseTranslate*, which implements *IMachineTranslation*, hooks the translation system to the user interface, and specifies two abstract methods *getPreferenceName*() and *translate*(…). To add a new MT system to OmegaT merely requires extending the *BaseTranslate* class, overriding the *getName*(),*getPreferenceName*() and *translate*(…) methods, and adding a new property to the Bundles.properties file to indicate, which menu the item should appear and how it should appear (MT_ENGINE_JOSHUA=Dari Legal). The Unified Modeling Language (UML) class diagram is shown in figure 4.

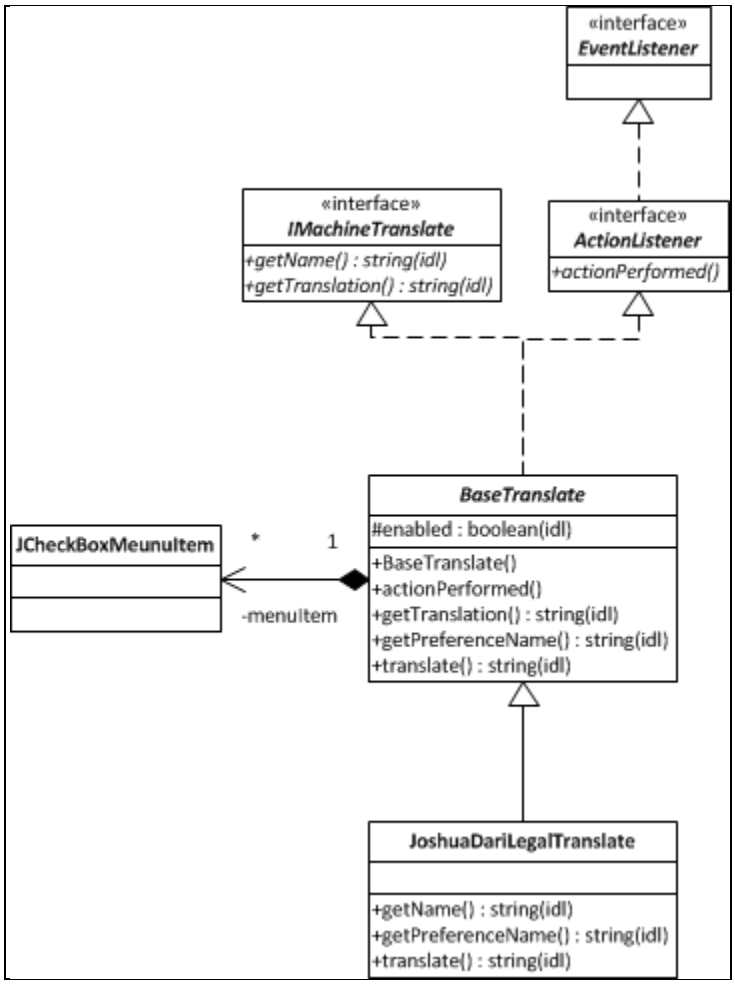

Figure 4. JoshuaDariLegalTranslate class diagram.

To use the system, the translator first enables the translation engine by selecting it from the Options menu (multiple translation engines can be active simultaneously, see figure 5). As the user processes each segment in the Editor panel on the left, the proposed Dari translation is shown in the Machine Translation panel on the right (figure 6).

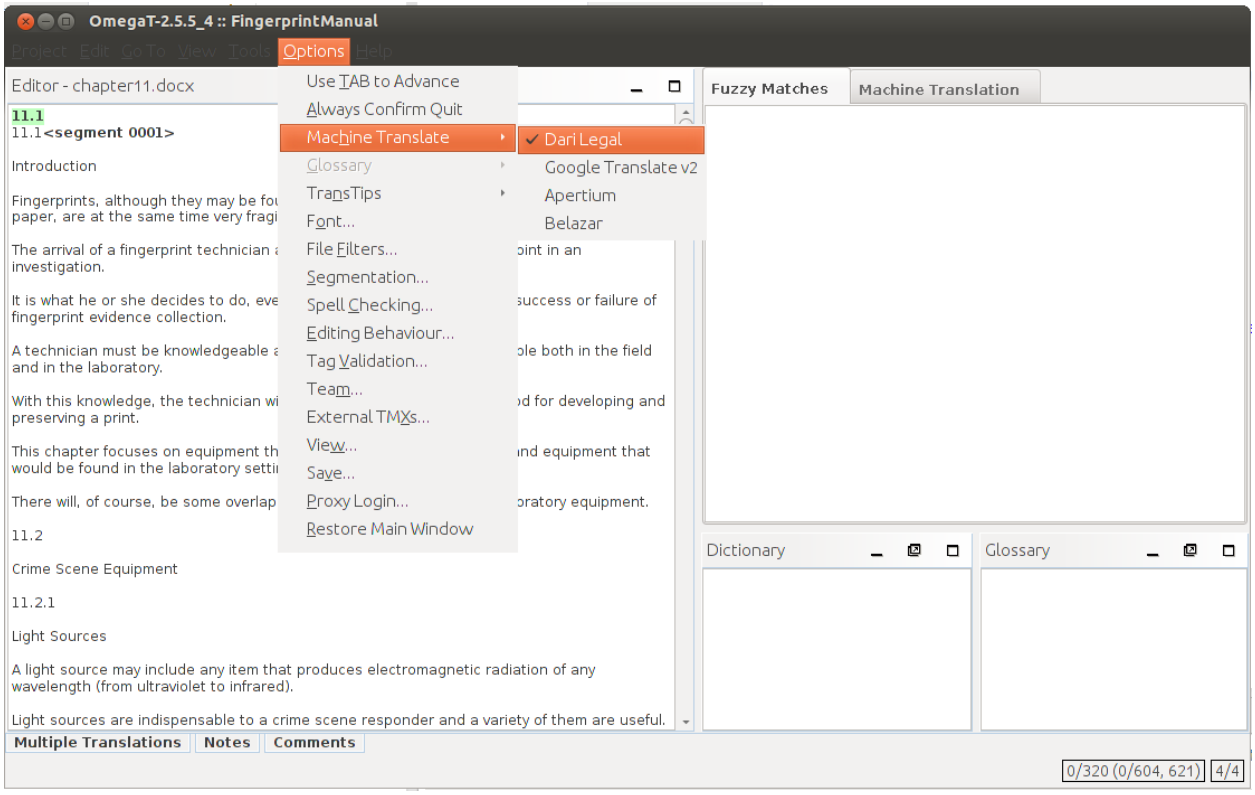
OmegaT-2.5.5_4 :: FingerprintManual
Project Edit GoTo View Tools Options Help
Editor - chapter11.docx
11.1
11.1<segment 0001>
Introduction
Fingerprints, although they may be fou
paper, are at the same time very fragi
The arrival of a fingerprint technician a
investigation.
It is what he or she decides to do, eve
fingerprint evidence collection.
A technician must be knowledgeable a
and in the laboratory.
With this knowledge, the technician wi
preserving a print.
This chapter focuses on equipment th
would be found in the laboratory setti
There will, of course, be some overlap
11.2
Crime Scene Equipment
11.2.1
Light Sources
A light source may include any item that produces electromagnetic radiation of any wavelength (from ultraviolet to infrared).
Light sources are indispensable to a crime scene responder and a variety of them are useful.
Use TAB to Advance
Always Confirm Quit
Machine Translate
Glossary
TransTips
Font...
File Filters...
Segmentation...
Spell Checking...
Editing Behaviour...
Tag Validation...
Team...
External TMXs...
View...
Save...
Proxy Login...
Restore Main Window
Dari Legal
Google Translate v2
Apertium
Belazar
Fuzzy Matches
Machine Translation
Dictionary
Glossary
Multiple Translations
Notes
Comments
0/320 (0/604, 621)
4/4

Figure 5. Enabling the Joshua Dari Legal SMT model from OmegaT.

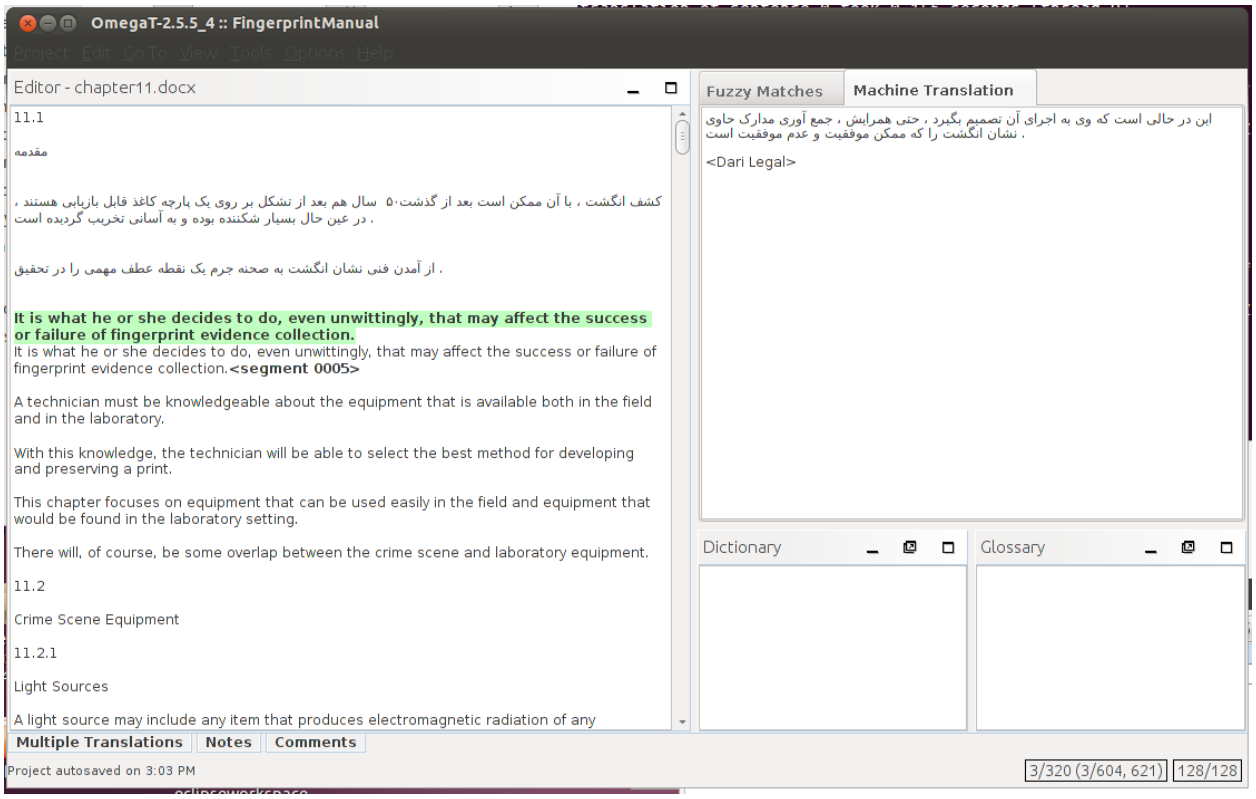
OmegaT-2.5.5_4 :: FingerprintManual
Project Edit GoTo View Tools Options Help
Editor - chapter11.docx
11.1
مقدمه
کشف انگشت ، یا آن ممکن است بعد از گذشت ۵۰ سال هم بعد از تشکل بر روی یک پارچه کاغذ قابل بازیابی هستند ، در عین حال بسیار شکننده بوده و به آسانی تخریب گردیده است .
از آمدن فنی نشان انگشت به صحنه جرم یک نقطه عطف مهمی را در تحقیق .
It is what he or she decides to do, even unwittingly, that may affect the success or failure of fingerprint evidence collection.
It is what he or she decides to do, even unwittingly, that may affect the success or failure of fingerprint evidence collection.<segment 0005>
A technician must be knowledgeable about the equipment that is available both in the field and in the laboratory.
With this knowledge, the technician will be able to select the best method for developing and preserving a print.
This chapter focuses on equipment that can be used easily in the field and equipment that would be found in the laboratory setting.
There will, of course, be some overlap between the crime scene and laboratory equipment.
11.2
Crime Scene Equipment
11.2.1
Light Sources
A light source may include any item that produces electromagnetic radiation of any
Multiple Translations
Notes
Comments
Project autosaved on 3:03 PM
Fuzzy Matches
Machine Translation
این در حالی است که وی به اجرای آن تصمیم بگیرد ، حتی همرایش ، جمع آوری مدارک حاوی نشان انگشت را که ممکن موفقیت و عدم موفقیت است .
<Dari Legal>
Dictionary
Glossary
3/320 (3/604, 621)
128/128

Figure 6. Using the Dari Legal SMT model in OmegaT.

6. Conclusion

In this report, we have described a human-in-the-loop Iterative Post-Editing Refinement algorithm for generating SMT models for highly specialized domains. Additionally, we described integrating these SMT models with the open-source Joshua decoder and exposing these models as Web services, making them available to any user or system with Internet access. Finally, we demonstrated how one such system might use these services by integrating them with the open-source CAT tool OmegaT. We are in the process of translating the National Institute of Justice's *Fingerprint Sourcebook* from English into Dari and will report on improvements in BLEU score once complete.

7. References

Ambati, V. *Active Learning and Crowdsourcing for Machine, CMU-SCS-11-020.* PhD Thesis, Pittsburgh: Carnegie Mellon University, 2012.

Ambati, V.; Vogel, S. Can Crowds Build Parallel Corpora for Machine Translation Systems? *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* . 2010, 62–65.

Ananthakrishnan, S.; Prasad, R.; Stallard, D.; Natarajan, P. Batch-Mode Semi-Supervised Active Learning for Statistical Machine Translation. *Computer Speech and Language* **2001**, *27*, 397–406.

Bakhshaei, S.; Khadivi, S. A Pool-Based Active Learning Method for Improving Farsi-English Machine Translation System. *6'th International Symposium on Telecommunications.* Tehran, 2012, 822–826.

Bertoldi, N.; Federico, M. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. *Proceedings of the 4th EACL Workshop on Statistical Machine Translation.* Athens: Association for Computational Linguistics, 2009, 182–189.

Callison-Burch, C. *Active Learning for Statistical Machine Translation.* PhD Thesis Proposal, Edinburgh University, 2003.

Callison-Burch, C.; Zaidan, O. Crowdsourcing Translation: Professional Quality from Non-Professionals. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.* Portland: Association for Computational Linguistics, 2011, 1220–1229.

Callison-Burch, C.; Zaidan, O. Feasibility of Human-in-the-Loop Minimum Error Rate Training. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.* Singapore: ACT and AFNLP, 2009, 52–61.

Dugast, L.; Senellart, J.; Koehn, P. Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. *Proceedings of the Second Workshop on Statistical Machine Translation.* Prague: Association for Computational Linguistics, 2007, 220–223.

Foster, G.; Goutte, C.; Kuhn, R. Discriminative InstanceWeighting for Domain Adaptation in Statistical Machine Translation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* . Massassachusetts, 2010, 451–459.

González-Rubio, J.; Ortiz-Martinez, D.; Casacuberta, F. An Active Learning Scenario for Interactive Machine Translation. *ICMI'11.* Alicante, Spain: ACM, 2011, 197–200.

Haffari, G.; Sarkar, A. Active Learning for Multilingual Statistical Machine Translation. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP.* Suntec, Singapore, 2009, 181–189.

Lagarda, A.-L.; Alabau, V.; Casacuberta, F.; Silv, R.; Diaz-de-Liano, E. Statisical Post-Editing of a Rule-Based Machine Translation System. *Proceedings of NAACL HLT* . Bouolder: Association for Computational Linguistics, 2009, 217–220.

Luhn, H. The Automatic Creation of Literature Abstracts. *IIBM J. Res. Devel.* **1958**, 159–165.

Mandal, A. et al. Efficient Data Selection for Machine Translation. *IEEE Workshop on Spoken Language Technology.* Goa, India: IEEE, 2008, 261–264.

Morgan, J. J. *Human in the Loop Machine Translation of Medical Terminology;* ARL-MR-0743; U.S. Army Research Laboratory: Adelphi, MD, 2010.

Robertson, S. E.; Jones, K. Sparck. Relevance Weighting of Search Terms. *J. Amer. Soc.* **1976**, *27* (3), 129–146.

Sennrich, R. Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.* Avignon, FR: Association for Computational Linguistics, 2012, 539–549.

Simard, M.; Ueffing, N.; Isabelle, P.; Kuhn, R. Rule-Based Translation With Statistical Post-Editing. *Proceedings of the Second Workshop on Statistical Machine Translation.* Prague: Association for Computational Linguistics, 2007, 203–206.

Terumasa, E. Rule-Based Machine Translation Combined with Statistical Post Editor fro Japanese to English Patent Translation. *MT Summit XI Workshop on Patent Translation.* Copenhagen, 2007.

Wang, W.; Machery, K.; Machery, .W.; Och, F.; Xu, P. Improved Domain Adaptation for Statistical Machine Translation. *Proceedings of the 2012 Association for Machine Trasnlation in the Americs.* San Diego, 2012.

Yu, C. T.; Salton, G. Precision Weighting—An Effective Automatic Indexing Method. *J. ACM* **1976**, *23* (1), 76–88.

List of Symbols, Abbreviations, and Acronyms

ARL	U.S. Army Research Laboratory
BLEU	Bilingual Evaluation Understudy
CAT	computer-assisted translation
CENTCOM	U.S. Central Command
MT	machine translation
OEF	Operation Enduring Freedom
OOV	out of vocabulary
REST	Representational State Transfer
SMT	statistical machine translation
TM	Translation Memory
UML	Unified Modeling Language

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	GOVT PRINTG OFC A MALHORTA
1 (PDF)	DIRECTOR US ARMY RESEARCH LAB IMAL HRA
1 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL
10 (PDS)	DIRECTOR US ARMY RESEARCH LAB RDRL-CII-B L TOKARCIK R WINKLER S METU
4 (PDFS)	RDRL-CII-T S LAROCCA J MORGAN M HOLLAND M VANNI
3 (PDFS)	RDRL-CII-C M THOMAS M MITTRICK J RICHARDSON
1 (PDF)	DIRECTOR NATIONAL SECURITY AGENCY CAMT M MARTINDALE
1 (PDF)	DIRECTOR ODNI FLPO J KLAVENS
1 (PDF)	DIRECTOR HQDA ODCS, G2 CSM M SANCHEZ
1 (PDF)	DIRECTOR CERDEC RDER-CPM-IM R SCHULZE
1 (PDF)	DIRECTOR INSCOM R RICHARDSON